Precise Generation of Conformational Ensembles for Intrinsically Disordered Proteins via Fine-tuned Diffusion Models

Junjie Zhu[1,#], Zhengxin Li[1,#], Zhuoqi Zheng[1], Bo Zhang[1], Bozitao Zhong[1], Jie Bai[1], Taifeng Wang[2,*], Ting Wei[1,*], Jianyi Yang[3,*], Hai-Feng Chen[1,*]

[1]State Key Laboratory of Microbial Metabolism, Department of Bioinformatics and Biostatistics, SJTU-Yale Joint Center for Biostatistics, National Experimental Teaching Center for Life Sciences and Biotechnology, School of Life Sciences and Biotechnology, Shanghai Jiao Tong University, 200240, Shanghai, China.

[2]BioMap,10 Beilun Industrial Park, Yongteng North Road, 100080, Beijing, China.

[3]MOE Frontiers Science Center for Nonlinear Expectations, Research Center for Mathematics and Interdisciplinary Sciences, Shandong University, 266237, Qingdao, China.

***Corresponding Author**

Email: haifengchen@sjtu.edu.cn; yangjy@sdu.edu.cn; weitinging@sjtu.edu.cn; taifeng@biomap.com

Notes

The authors declare that there is no conflict of interest.

Author Contributions

[#]These authors contributed equally to this work.

Abstract

Intrinsically disordered proteins (IDPs) play pivotal roles in various biological functions and are closely linked to many human diseases including cancer, diabetes and Alzheimer disease. Structural investigations of IDPs typically involve a combination of molecular dynamics (MD) simulations and experimental data to correct for intrinsic biases in simulation methods. However, these simulations are hindered by their high computational cost and a scarcity of experimental data, severely limiting their applicability. Despite the recent advancements in structure prediction for structured proteins, understanding the conformational properties of IDPs remains challenging partly due to the poor conservation of disordered protein sequences and limited experimental characterization. Here, we introduce IDPFold, a method capable of generating conformational ensembles for IDPs directly from their sequences using fine-tuned diffusion models. IDPFold bypasses the need for Multiple Sequence Alignments (MSA) or experimental data, achieving accurate predictions of ensemble properties across numerous IDPs. By sampling conformations at the backbone level, IDPFold provides more detailed structural features and more precise property estimation compared to other state-of-the-art methods. IDPFold is ready to be used in the elucidate the sequence-disorder-function paradigm of IDPs.

Introduction

Intrinsically disordered proteins (IDPs) constitute a category of proteins with unstable structures under physiological conditions[1]. These proteins, accounting for over 40% of eukaryotic proteomes[1], are involved in various biological functions including signal transduction, molecular recognition, and cell cycle regulation[2–4]. IDPs are also closely associated with various significant diseases, such as cancer, Parkinson's disease, and acquired immunodeficiency syndrome (AIDS)[5–7]. Unlike structured proteins that possess one or a few stable conformations, IDPs exhibit transitions between multiple conformations with very low energy barriers, constantly fluctuating within a broad ensemble of structures under physiological conditions[8]. Consequently, deciphering the conformational ensemble of IDPs poses a significant challenge for experimental methods such as X-ray, cryo-electron microscopy and NMR[9–12].

Currently, molecular dynamics (MD) simulation is the most commonly used and effective tool for sampling conformational ensembles[13–15]. By iteratively sampling target molecular system based on first principles, an estimation of conformational ensemble is obtained from simulation. However, MD simulation incurs high sampling costs, and ensuring thorough sampling of the ensemble is challenging. Additionally, commonly used force fields for IDPs, such as ff03CMAP[16], ff14IDPs[17], ESFF1[18], still exhibit considerable errors in estimating local and global properties of IDPs during simulations[19,20]. Although we can adjust MD simulations to better align with the characteristics of IDPs using experimental data, only a few of these proteins have been experimentally characterized[21]. These limitations constrain the application of MD in predicting the conformational ensemble of IDPs.

On the other hand, numerous deep learning methods have been developed and widely applied in the field of structure prediction for structured protein, such as AlphaFold2 and ESMFold[22,23]. Simultaneously, in protein design tasks, various generative models, such as RFdiffusion and Chroma, have been utilized for generating diverse protein backbones[24,25]. Therefore, it is natural to consider whether deep learning methods can be employed for rapid and accurate prediction of conformational

ensembles of IDPs. However, the structures associated with IDPs are currently very sparse. The PDB database contains $> 220{,}000$ structural entries[26], while the Protein Ensemble Database only includes 553 protein ensemble data[27]. Moreover, IDPs often lack high-quality multiple sequence alignment (MSA) data. MSA has been proven effective only in functional studies concerning folded or bound states of IDPs, and provides little assistance in predicting more disordered conformations[28]. These challenges make it highly daunting to use deep learning for robustly predicting the conformational ensembles of IDPs.

We introduce IDPFold here for predicting IDP dynamics directly from sequences based on a generative deep learning model. IDPFold utilized a protein language model to extract sequence information and further fed it into a structure generation module, enabling MSA-free conformation generation. To address the issue of insufficient data, we employed a hybrid dataset comprising crystal structures, NMR structures and MD trajectories for training IDPFold. The experimental structures enable the model to learn basic protein characteristics, while MD trajectories provide sufficient IDP structural data, ensuring accurate sampling on IDP systems. IDPFold generates IDP conformational ensembles at backbone level that is in better agreement with experimental observations than other state-of-the-art methods. IDPFold is able to sample conformational transitions from structured to disordered states, providing insights for studying the correlation between structures and functions of IDPs.

Results

Overview of IDPFold

IDPFold employs a conditional diffusion model framework for generating protein conformational ensembles from sequences (Figure 1A). This framework involves a forward diffusion process where noise is gradually added to real protein structures, and a reverse diffusion process where a deep learning network is used for denoising. By

integrating specific protein sequence features into the model during the reverse diffusion process, we can generate conformational ensembles for specific proteins using this architecture. The denoising network consists of 4 DenoisingIPA Blocks, which are similar to the structural modules in AlphaFold2. This network is to capture the chain-like structure within proteins and the rotational/translational state of each residue. For more details please refer to Materials and Methods section.

We first evaluated the backbone bond lengths, bond angles, and dihedral angle distributions of the conformations generated by IDPFold. Since each residue was treated as a rigid body during the conformation generation process and parameterized using ideal relative position, we focused primarily on inter-residue features. The conformations generated by IDPFold closely matched the distributions of bond lengths and bond angles observed in MD trajectories, indicating that the model learned the chain-like arrangement between adjacent residues and accurately captured the spacing between neighboring residues (Figure 1B, Figure S1). Additionally, the distribution of Ω angles in the conformations generated by the model resembled that observed in MD trajectories, with a peak around 180°, suggesting that the local peptide bond plane conformations generated by the model are stable and align with general protein characteristics. Moreover, the Ramachandran plot of $\varphi - \psi$ angles shows that the backbone dihedral angles of generated conformations predominantly fall within reasonable regions. The density estimates across various regions closely approximate those observed in MD trajectories (Figure 1C). Specifically, with IDPFold fine-tuned on IDRome data, there is a notable improvement on the probability of the ppII region in the upper left corner compared to the untuned version, showing a distribution closer to MD trajectories.[13] This indicates that the fine-tuning enables the model to generate more disordered structures, capturing the intrinsic biases of target proteins more precisely (Figure S2).

Next, we used IDPFold to predict the conformational ensembles on four protein systems containing significant disordered regions, demonstrating the differences between the model-generated conformations and those observed in MD trajectories

(Figure 1D). The model-estimated local secondary structures and overall shape closely resemble those generated by MD, with stable conformational estimation for structured regions. As a comparison, the generated conformations exhibit greater diversity in disordered regions, aligning with the highly flexible physiological characteristics of these regions.

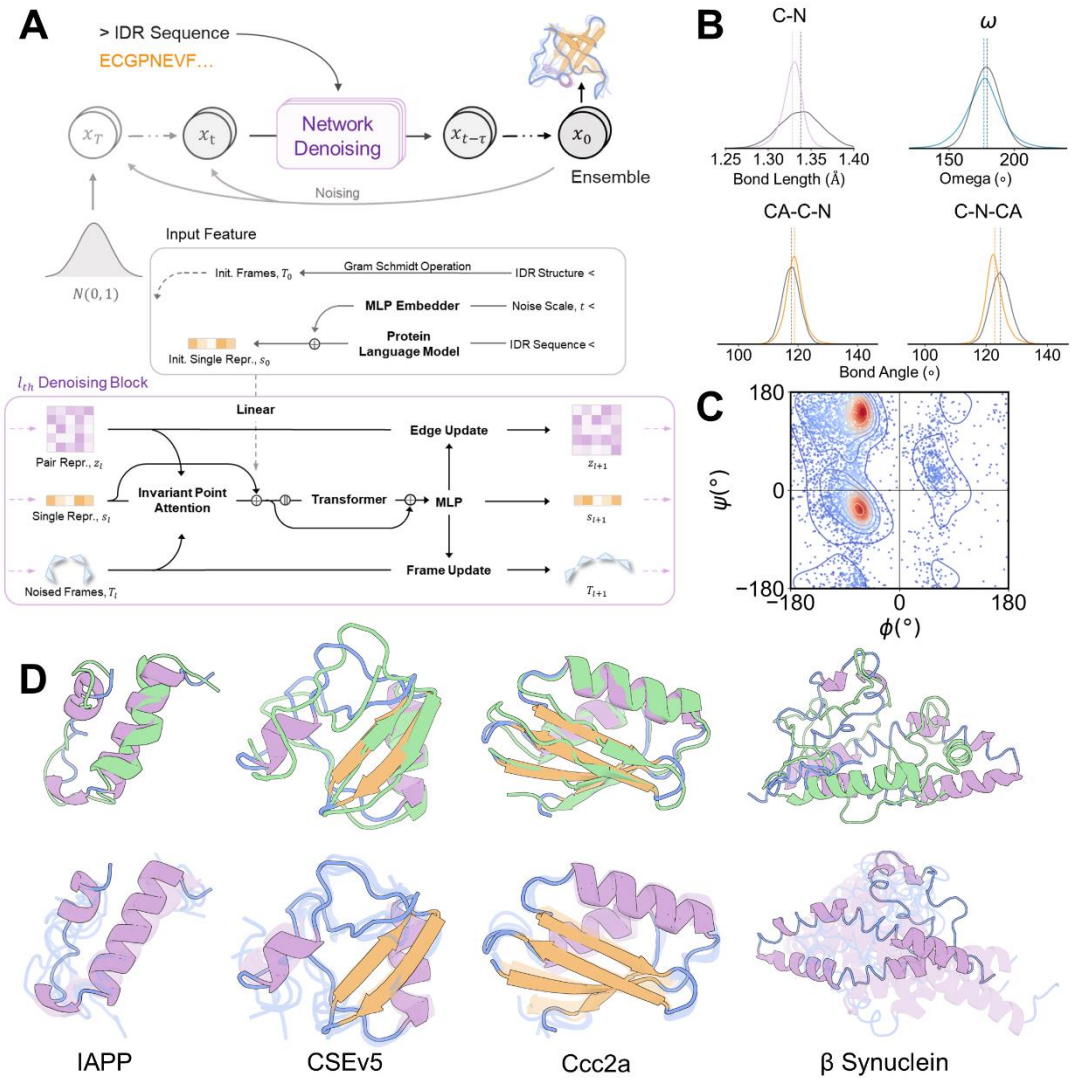

Figure 1. Overview of IDPFold. (A) Model architecture and framework for generating IDP conformation ensembles. (B) Intra-residue bond lengths, bond angles and dihedral ω of IDPFold-generated conformations compared to MD trajectories (grey). (C) Ramachandran plot, scatters are from generated samples and density lines are from MD trajectories. (D) IDPFold-generated conformations on several systems and similar

structures (green) from MD trajectories. Generated conformations are colored according to their secondary structures.

Evaluate the fidelity of IDPFold on experimental data

To assess the fidelity of the dynamics conformational ensembles predicted by IDPFold, we collected experimental physical quantities from 27 IDP systems of varying lengths (Table S1) and performed 1 μs all-atom molecular dynamics simulation for each system. All simulations were utilized with the ESFF1 force field along with the OPC3-B solvent model[18,29], which are specifically targeted for disordered proteins. Convergence of the simulation trajectories was confirmed.

We first evaluated the local features of generated ensembles using chemical shifts as the target physical quantity to measure how generated ensembles differ from experimental observations (Figure 2A). Specifically, we computed C_α secondary chemical shifts of generated conformations and the mean absolute error (MAE) between ensemble average and experimental values. C_α secondary chemical shifts can reflect the secondary structure propensity of a protein system, where positive values indicate a higher helical propensity in certain region, negative values indicate a tendency to form strands, and values close to zero indicate disordered regions. The ensembles generated by fine-tuned IDPFold are close to experimental observations, showing an average MAE of 0.48. Pairwise t-tests revealed significant improvement over MD simulations in fitting chemical shifts ($p = 0.03$), whereas pre-trained IDPFold performed worse than MD in fitting chemical shifts.

We next assessed ensemble average radius of gyration (Rg) of the generated ensembles (Figure 2B). Rg is a critical physical quantity that reflects the global characteristics of IDPs, with its magnitude positively correlating with overall protein looseness. Rg is often used to describe and distinguish IDPs from structured proteins[29]. We calculated the average error and standard deviation relative to experimental Rg values for conformational ensembles generated by IDPFold. Those generated by

IDPFold pretrained on experimental structures significantly underestimated Rg for most IDPs (average error of -22%), indicating a clear bias towards structured protein in prediction. This observation aligns with the high proportion of structured regions in experimental structures, lacking representation of natural disordered regions. However, the model significantly improved its Rg estimation after fine-tuning on MD trajectories, with an average error of only -7%. Paired t-test demonstrates a significant improvement over all-atom simulation (average error of -19%) with $p = 3.57 \times 10^{-4}$.

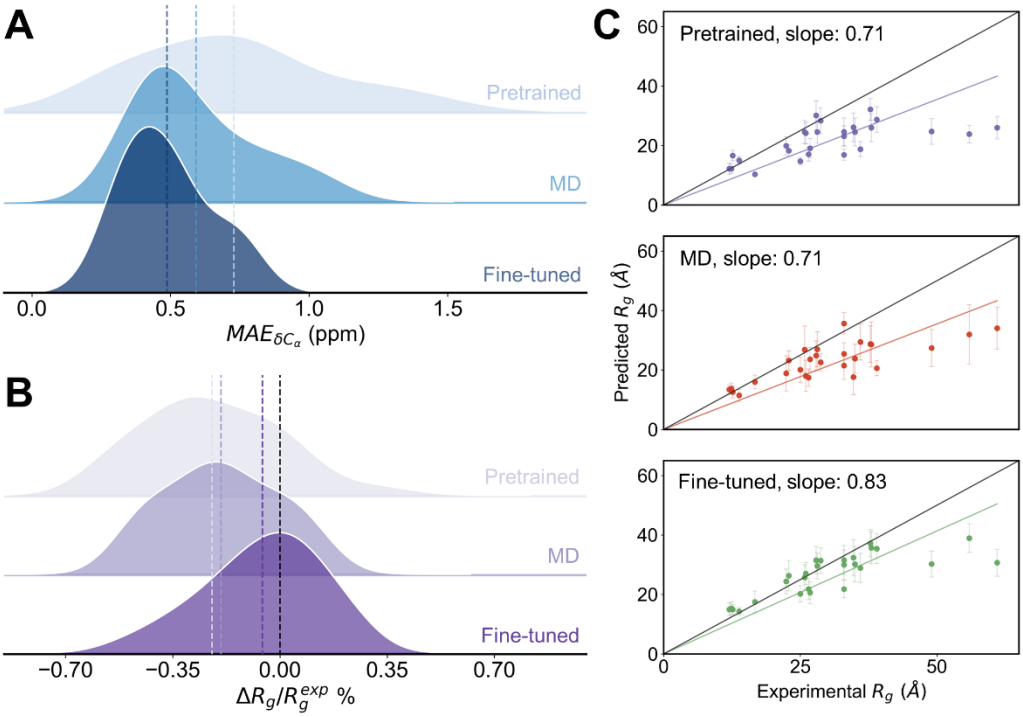

Figure 2. Fidelity of IDPFold-generated dynamics conformational ensembles evaluated on experimental data. (A) Ensemble average Rg error, (B) Mean absolute error on ensemble average C_α chemical shift of MD trajectories (middle), pretrained (upper) and fine-tuned (lower) IDPFold against experimental values. (C) Regression plot between experimental Rg values and predicted Rg values for MD simulation and IDPFold. The intercept was set to zero.

IDPFold not only accurately captured the ensemble average Rg of IDPs but also exhibited diversity in generated conformations comparable to MD simulations (Figure

S3). Particularly, for IDP systems with a higher structured tendency, such as ACTR and drkN SH3, the pretrained IDPFold exhibited minimal variance and mean Rg in generated structures, indicating limitied diversity and excessive compactness. As a comparison, the fine-tuned model showcased average Rg values closer to experimental observations with deviations comparable to MD trajectories ($p = 0.22$ in paired t-test), suggesting both a higher diversity in generated conformations and better alignment with experimental states.

Finally, we assessed the performance of IDPFold across various test systems. For most cases, IDPFold estimated the conformational ensemble chemical shifts more accurately than MD simulation (Figure 3A, S5). IDPFold generally provided higher Rg estimates for these proteins compared to all-atom MD simulations. However, in terms of the range of Rg distributions, the diversity of conformations generated by IDPFold was comparable to that of MD (Figure 3B). Additionally, we randomly selected three conformations from each system for visualization. These conformations exhibited significant diversity in disordered regions while correctly maintaining stability in structured regions such as the helices in PaaA2 (Figure 3C).

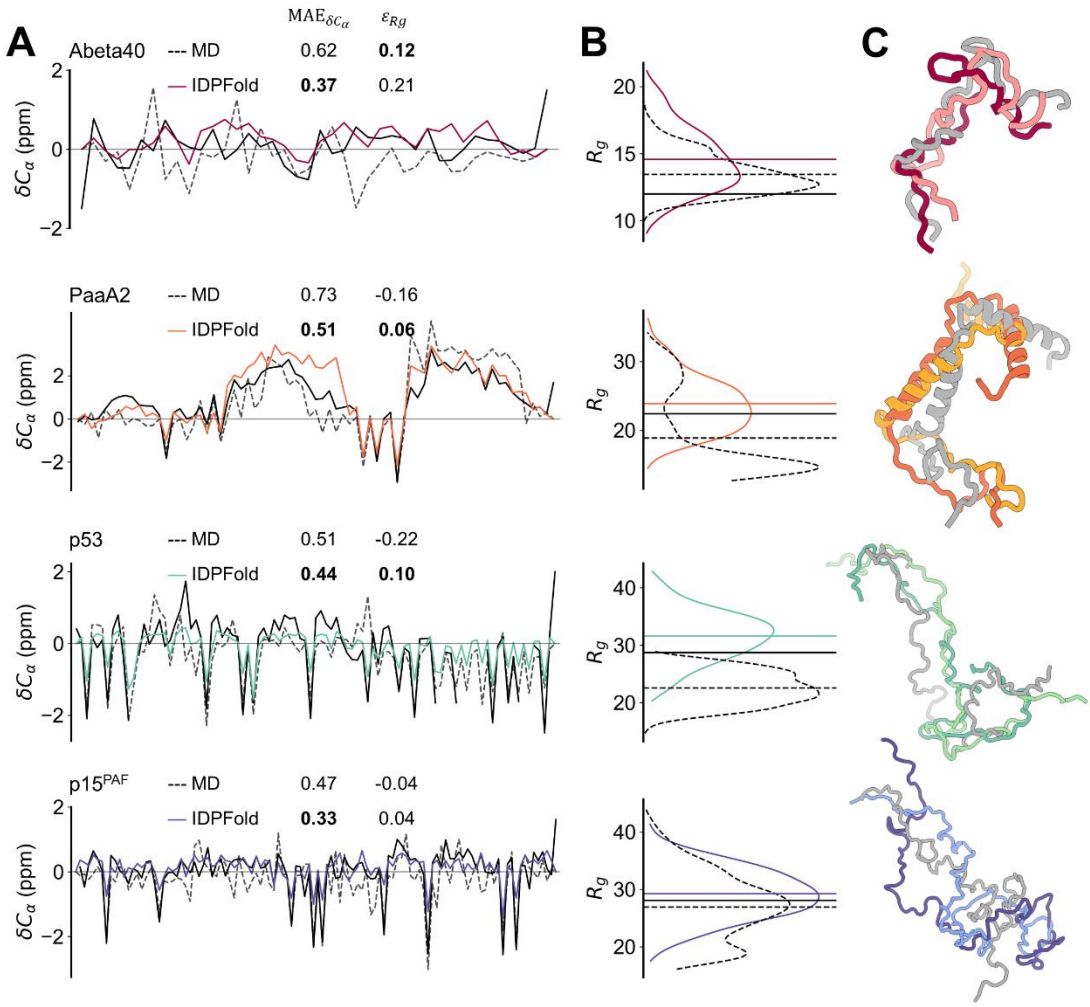

Figure 3. Several cases of IDPFold generated conformational ensembles. (A) C_α chemical shifts, (B) Rg distributions of generated ensembles (colored) and MD trajectories (dashed black). Experimental values are plotted as solid black lines. Average error on chemical shift and Rg values are listed, where bold values denotes better performance. (C) Conformations generated by IDPFold.

Establish a benchmark of conformation sampling methods on IDPs

We established a benchmark on the 27 IDP systems and compared the performance of IDPFold with conventional MD and current state-of-the-art deep learning-based ensemble prediction methods. We computed the validity for each conformational ensemble as well as the fidelity composed of ε_{Rg} and $MAE_{\delta C_\alpha}$, and recorded the time

required for sampling ensembles by these methods (Table 1). A higher validity indicates a higher proportion of reasonable distances between adjacent C_α atoms in conformations, while the value close to 0 is preferred for both ε_{Rg} and $MAE_{\delta C_\alpha}$, indicating better alignment with experimental observations. Together, they represent the agreement of the predicted ensemble to experimental observations. Specific definitions of these indicators can be found in the Materials and Methods section.

In terms of validity, all deep learning-based methods performed well, where AF-cluster achieving the highest accuracy in estimating neighboring C_α distances, with a validity of 0.99. AF-cluster utilizes AlphaFold2 to generate multiple conformations by clustering the MSA information of input sequence, contributing to high-quality generated structures. However, the test system RS1 lacked MSA information. For other IDP systems, the number of conformations produced by AF-cluster was also limited by the availability of MSA data, making AF-cluster less versatile for IDP applications[30].

IDPFold pretrained on experimental structures exhibited similar validity to AF-cluster, with the fine-tuned version showed slightly lower validity compared to the pretrained versions. We suggest the reasons as highly flexible IDRs often undergo significant fluctuations in atomic distances during motion in MD trajectories, with the range of local structural changes far exceeding that of experimental structures. Therefore, after fine-tuning, IDPFold tended to generate protein local structures that fluctuated more significantly.

In terms of fidelity, the fine-tuned IDPFold performed the best across both experimental metrics (Figures S4-S5). For AlphaFlow, we tested all eight versions and found that AlphaFlow-PDB-base performed the best among them (Table S2)[31]. Regarding runtime, the two methods based on the AlphaFold2 architecture, AF-cluster and AlphaFlow, are very slow. In contrast, IDPFold, having removed the MSA processing module and designed a more lightweight structure module, took an average of only about 20 minutes to generate 300 conformations. Meanwhile, although idpSAM-generated structures were overly compact and deviated significantly from experimental values in terms of Rg, its inference speed was very fast, averaging only

0.63 minutes to sample protein conformations[32]. The model used in idpGAN had the fewest parameters.[33]. Though it did not perform so well as fine-tuned IDPFold in both validity and fidelity, it required only 0.05 seconds on average to sample conformations. Overall, the fine-tuned IDPFold exhibited the best fit to experimental observations of IDP ensembles, and its conformation generation speed was acceptable, making it suitable for IDP ensemble prediction tasks. The performance of each method on ensemble average Rg is shown in Figures S6-S9, and the performance on chemical shifts is shown in Figures S10-S13.

Table 1. Benchmark on IDPFold and other methods. Bold values denote the best.

Methods	**Validity (↑)**	**Fidelity**		**Time**
		ε_{Rg}	$MAE_{\delta C_\alpha}$ (↓)	
MD $1\mu s$	-	-0.19	0.59	~200 h
idpGAN	0.91	-0.12	0.56	**0.05 sec**
idpSAM	0.95	-0.55	0.79	0.63 min
AlphaFlow (best)	0.97	-0.24	0.54	78.20 min
AF-cluster	**0.99**	-0.12	0.57	~280 min
IDPFold, Pretrained	0.98	-0.22	0.73	20.35 min
IDPFold, Fine-tuned	0.95	$\mathbf{-0.07}$	**0.48**	

We further selected 58 proteins from the IDRome dataset that were not included in the IDPFold training set[34,35]. We compared the conformations generated by IDPFold with the trajectories obtained from coarse-grained simulations based on their Rg differences. The average Rg error estimated by IDPFold for these systems was -0.04, which is comparable to the error level reported for coarse-grained simulations in IDRome (Figure 4A)[21]. Additionally, the range of Rg distributions sampled by IDPFold for most systems was similar to that of coarse-grained simulations, further demonstrating that IDPFold sampled diverse IDP conformations (Figure 4B, S14).

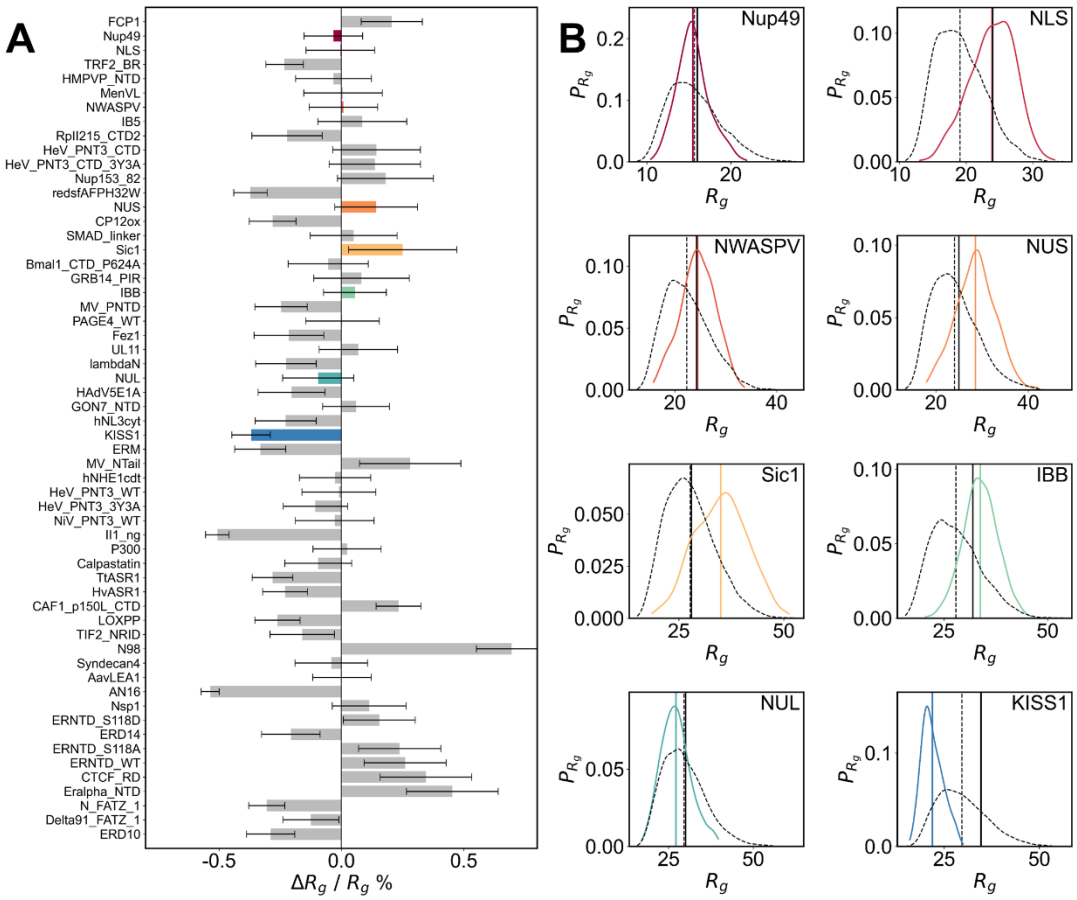

Figure 4. Comparing the generated ensembles of IDPFold and CALVADOS 2[35]. (A) Relative error between IDPFold predicted and experimental radii of gyration. (B) Rg distributions of IDPFold generated ensembles (colored) and CALVADOS simulations (black dashed). Experimental values are plotted as solid black lines.

IDPFold captures conformational changes of marginally disordered proteins

IDPs that lie close to the empirical boundary separating IDPs and folded proteins in Uversky's charge–hydropathy plot may behave as "marginal IDPs"[36]. Marginal IDPs are sensitive to changes in their environment or sequence, leading to a wide range of conformation change under different conditions. They range from being mostly folded with some flexible linkers to disordered structures that fold upon binding for molecular recognition or even fully disordered yet functional forms[37]. Here, we selected two

marginal IDPs, CT10-Regulated Kinase (Crk1) and Huntingtin-interacting protein K (HYPK), and used IDPFold to sample their structural transitions between ordered and disordered conformations.

Crk1 (UniProt ID: P46108-1) is a human cancer-related signaling adaptor protein. It primarily consists of an IDR linker of 13 residues connecting the SH2 and SH3 domains, with an additional 100-residue region following the SH3 domain that remains unresolved in X-ray structures (Figure 5A)[38,39]. We generated full-length Crk1 conformations using IDPFold and applied PCA for dimensionality reduction. PCA is useful for analyzing and visualizing the similarity between conformations.

The resulting conformational ensemble mainly exhibited two major types: the highly disordered Peak 1 and the relatively structured Peak 2 (Figure 5B). Additionally, we presented the highest-scoring conformation predicted by AlphaFold3[40]. We found that IDPFold was capable of sampling extensive conformational variations for both individual domains and the IDR linker, producing structures similar to experimental observations. In the relatively structured conformations generated by IDPFold, the local secondary structure estimations were quite accurate. We further computed distance maps for the experimental structure and three predicted structures, discovering that AlphaFold3 predicted the IDR linker region in a turn-like conformation, bringing the SH2 and SH3 domains unusually close to each other (Figure 5C). As a comparison, IDPFold predominantly sampled more extended linker conformations, resulting in a greater distance between the two domains, aligning better with the experimentally observed characteristics in linker. Additionally, AlphaFold3 predicted a long, nearly circular loop for the unresolved region, ending in a compact area composed of beta strands and loops. A low pLDDT score was predicted for the long loop region, indicating it was indeed disordered but with unreliable local conformations. While AlphaFold3 accurately predicted the IDR regions, the reasonableness of the IDR in its predicted structure is debatable. Conversely, IDPFold provided similar secondary structure estimations and also predicted loop region conformations that were more physically plausible.

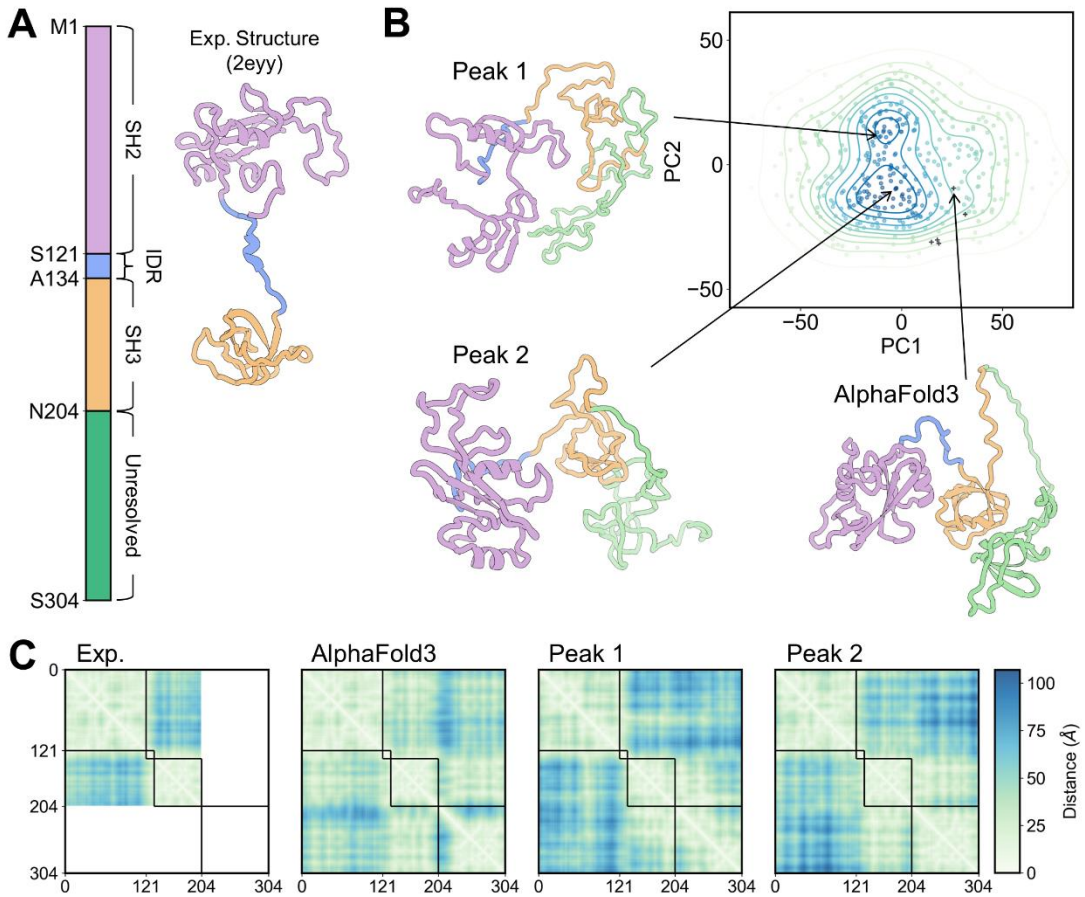

Figure 5. IDPFold-generated conformational ensemble of Crk1. (A) The experimental structure and domain information of Crk1. (B) Mapping IDPFold-generated conformations and AlphaFold3-predicted structures onto the top two principal component space. (C) Distance map of the experimental structure, AlphaFold3-predicted structure, and two peak structures generated by IDPFold.

HYPK (UniProt ID: Q9NX55) is a translation regulator essential in Huntington's disease. A previous study using FRET reported that mutating a group of acidic amino acids to basic ones in its sequence increases the ensemble average end-to-end distance, indicating that changes in charge made the protein more disordered[37,41]. Here, we used IDPFold to generate conformational ensembles for both the wild-type and mutant HYPK (E16K/E20K/E51K/E101K), and examined the similarities and differences between the two ensembles in the Rg and end-to-end distance plane (Figure 6A). We

found that for both sequences, IDPFold sampled a wide range of conformations from structured to disordered. The ensemble generated for the mutant HYPK showed higher Rg and end-to-end distance distributions and means compared to the wild-type, consistent with the previous FRET observations.

We visualized the high-scoring conformations predicted by AlphaFold3 (Figure 6B) and the structured and disordered conformations in the generated ensemble (Figures 6C, 6D). The predicted structures were divided into two parts: the long N-terminal loop and the relatively stable three-helix region from K91 to N129 at the C-terminal. In terms of chemical shifts, the conformations generated by IDPFold were similar to those predicted by AlphaFold3, both showing very stable helices. For the helices near the loop region, IDPFold tended to predict a transition to disordered structures in the generated disordered conformations, while maintaining the third helix well in the structured conformations (Figure 6E). In the long N-terminal loop region, both IDPFold-generated conformations and AlphaFold3-predicted structures accurately showed a disordered tendency. IDPFold sampled a wide range of conformations from partially helical and globular structures to fully disordered chain-like structures for this loop. From the distance maps, Rg, and end-to-end distances, the AlphaFold3-predicted structure was intermediate between the two types of conformations generated by IDPFold and IDPFold sampled ensemble could cover conformations similar to those predicted by AlphaFold3 (Figure 6F).

In summary, IDPFold is capable of sampling large-scale conformational transitions in marginal IDPs from structured to disordered conformations and can reasonably model the impact of charge changes on disorder tendency in sequences. This makes it suitable for studying function-related conformational changes in IDPs under physiological conditions.

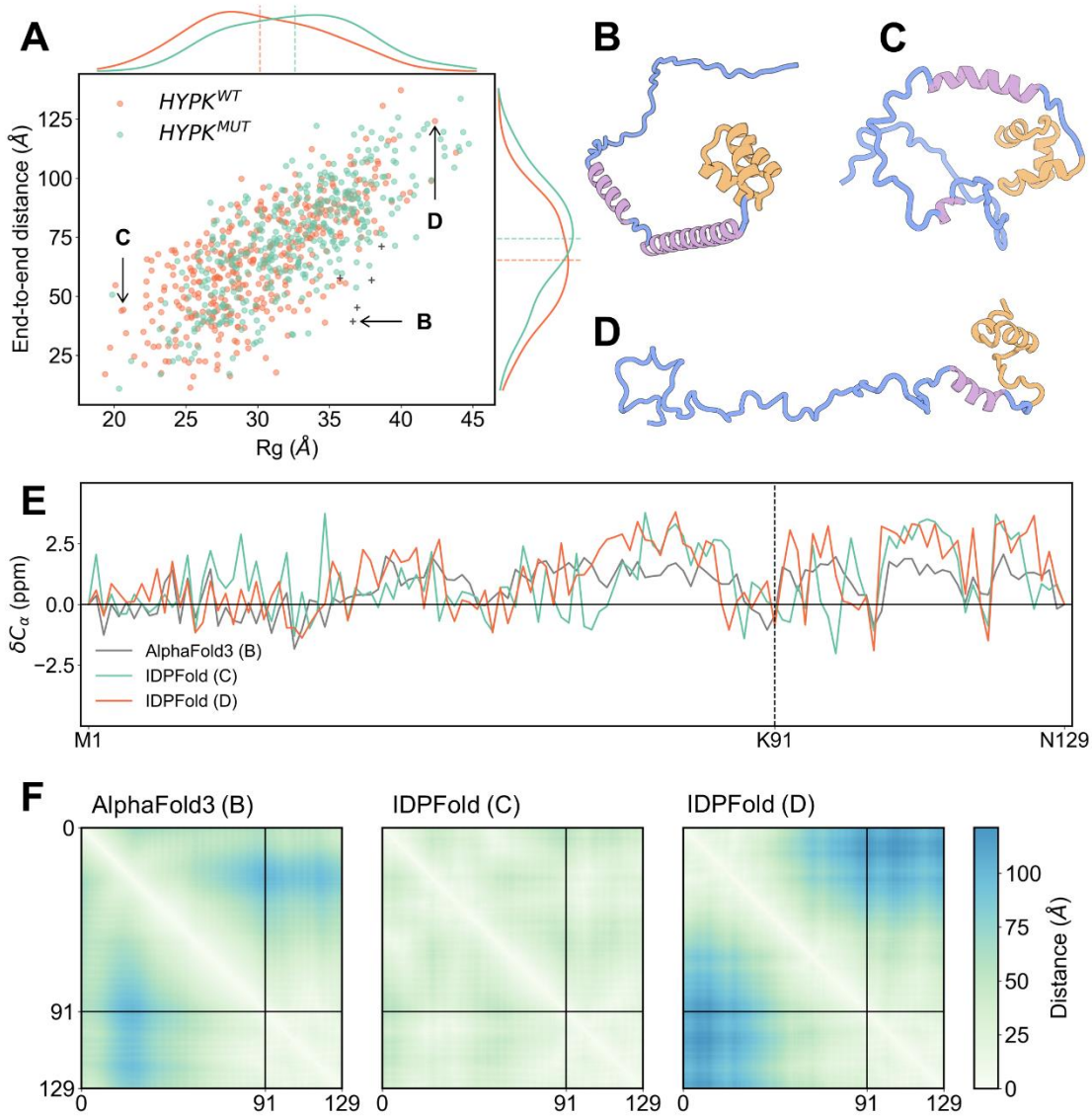

Figure 6. IDPFold-generated conformational ensembles of wild type and mutant HYPK. (A) Distribution of generated ensembles on Rg and end-to-end distance. (B) AlphaFold3-predicted structure. (C) and (D) IDPFold-generated structures, corresponding to the annotations in (A). (E) C_α chemical shifts. (F) Distance maps of the AlphaFold3-predicted structure and IDPFold-generated structures.

Discussion

In this study, we developed IDPFold, a deep learning-based tool for predicting IDP dynamics conformation ensembles. IDPFold adopts a conditional diffusion model architecture to perform an end-to-end conformational generation, integrating protein

language model for sequence feature extraction and DenoisingIPA module for conformation denoising. Through a two-stage training strategy on experimental data and MD trajectories of IDPs respectively, IDPFold efficiently and accurately samples IDP dynamics conformational ensembles. Furthermore, we established a benchmark of conformation sampling methods on 27 protein systems that contain IDRs. IDPFold precisely capture the overall compactness and local secondary structure features of IDPs, with predicted ensembles exhibiting features with values closer to experimental observations compared to both existing MD-based and deep-learning methods. Additionally, IDPFold can sample conformational changes from structured to disordered states in proteins, demonstrating a wide sampling range and high efficiency in ensemble sampling. This capability provides important insights for studying the conformational changes and functions of IDPs.

IDPFold achieves accuracy levels comparable to, or even higher than, traditional MD simulations in estimating IDP conformational ensembles with its sampling process not restricted by energy barriers. This is not only essential for studying IDP conformations but also holds tremendous promise for dynamic proteins like allosteric proteins and enzymes, which undergo large-scale conformational changes during functional processes. Although this work focused on training the model for IDPs and direct prediction performance on allosteric protein ensembles might not be optimal, the methodology would likely to be transferred on diverse categories of proteins, enabling more accurate estimations of conformational transitions in these functionally important proteins (Figure S15).

Although IDPFold demonstrates robustness and precision in predicting conformational ensembles, its inference time is relatively long compared to some of the existed deep learning methods with an average time of about 20 minutes to sample an entire system. While this speed is much faster than conventional MD simulations, there is room for further improvement on model efficiency. To achieve higher sampling accuracy and more precisely protein backbone features, we employed a more complex network architecture and longer diffusion steps which might well sacrifice sampling

efficiency. To enhance the inference efficiency of the model, optimizations can be explored such as reducing the number of diffusion steps in the diffusion model and optimizing transformer components by replacing them with architectures that are more efficient in terms of time and space. Additionally, in this study, we only trained and tested IDPFold on single-chain proteins and did not evaluate it on protein complexes or multi-chain proteins. In future work, we aim to extend the model to these areas to efficiently sample and interpret how molecules dynamically interact with each other.

Conformational ensemble prediction represents a significant frontier in protein structure research following static structure prediction. Accurate characterization of protein conformational ensembles can help us understand the dynamics of proteins and crucial conformational changes they undergo when binding to substrates or undergoing biochemical reactions under physiological conditions. Although deep learning-based tools are not yet widely adopted and most dynamic analyses still relying on MD simulations, our work suggest that properties predicted using deep learning methods might offer higher accuracy than MD simulations, potentially serving as complementary or even alternative approaches to MD simulations. However, current deep learning-based conformational ensemble prediction methods still face several challenges, such as inaccurate estimation of conformational free energies, less robustness compared to traditional methods, and the inability to capture temporal autocorrelation between conformations (i.e., inability to capture dynamic features).

Several studies have indicated that the single-point energy estimation of molecular force fields can be used to train diffusion models, allowing models to learn more comprehensively about conformational space in cases where structural data are scarce[42,43]. This training approach, which utilizes force field energies rather than structural data, enables more accurate capture of conformational space distribution and greater robustness. However, this training strategy also makes the model sensitive to empirical parameters of the force field, demanding high precision in molecular force field accuracy. Additionally, there are works based on flow models or score-matching models aiming to fit the dynamic characteristics of MD trajectories, which might be a

solution to the current inability of generative models to capture temporal autocorrelation between conformations[44].

Apart from the challenges of sparse training data and limited model representation capabilities, there is another significant issue in conformational ensemble prediction tasks: the scarcity and lack of uniformity in evaluation metrics. While we collected a set of experimental observations as a gold standard for this study, a vast number of protein systems lack experimental annotations, making it challenging to assess the generative performance of computational models on these systems. Therefore, we believe that constructing confidence metrics for conformational ensemble generation akin to pLDDT for protein structure prediction is also promising for evaluating and improving protein conformational ensemble models.

Materials and Methods

Datasets

The data for training IDPFold consists of three components: high-quality crystal structures collected from the PDB dataset, conformations ensembles resolved by NMR and molecular dynamics (MD) simulation dataset of IDPs obtained through extensive back-mapping and energy minimization on coarse-grained trajectories.

For crystal structures, we referenced trRosetta and gathered a total of 15,051 X-ray resolved structures with resolutions ≤ 2.5Å and sequence redundancy $\leq 30\%$[45]. These structures provide fundamental protein characteristics for the model, such as bond lengths, bond angle distributions and chirality of C_α. However, crystal structures typically depict stable conformations of proteins, i.e., structured conformations. Training the model solely on crystal structures would significantly underestimate the disorder tendency of IDPs and result in low diversity in conformation generation for individual protein sequences. Therefore, we additionally collected 12,339 NMR resolved protein conformational ensembles from the PDB. We filtered these NMR

entries based on 30% sequence similarity both internally and against crystal structures, resulting in 539 systems comprising a total of 10,454 structures. By blending crystal structures with NMR ensembles, we obtained a combined total of 25,495 experimental structures for the initial phase of IDPFold training. These structures were well-defined and exhibited a higher structured tendency, ensuring that the model captures the local physical characteristics of proteins effectively.

For the IDP trajectory data, we obtained large-scale coarse-grained simulation data from IDRome that recorded most IDRs in human proteome[21]. These coarse-grained simulations accurately capture global features of IDPs (and IDRs), such as average radius of gyration (Rg) of conformational ensembles, but they only retain C_α atoms, whose resolutions fail to meet the requirement for model training. Therefore, we selected all systems with lengths exceeding 256 residues, totaling 3,880 systems. Choosing larger systems is primarily because these long disordered segments encompass most sequence and structural characteristics found in smaller systems, and training the model on larger systems can help enhance its generalization ability. For these systems, we first used pdbfixer to restore structures from coarse-grained to all-atom and then performed 100 steps of energy minimization with ff14SB[46] force field for all protein structures. Evaluation showed that this optimization process effectively eliminated unreasonable conformations in the back-mapped structures (Figure S1). A total of 77,600 optimized all-atom conformations was used for the second phase of IDPFold training, these conformations globally exhibit more disordered and can aid the model in learning distinctive conformational features of natural IDPs.

Furthermore, data used to assess the performance of IDPFold generation included 27 IDP systems that are fully described by experiments and are not present in the training set. We conducted $1\mu s$ all-atom MD simulations for these systems with the IDP-specific force field ESFF1 and solvent model OPC3-B[18,29]. These all-atom simulation trajectories described the dynamics and thermodynamic characteristics of IDPs, which are suitable for evaluating the quality of IDPFold-generated conformational ensembles.

Formulation of IDPFold

Diffusion Modeling on Protein Structure

To enable IDPFold to capture Boltzmann distribution of protein conformations at equilibrium state, we employ Score-Based Generative Modeling (SGM) to learn the probability distribution from protein structure data. SGM can be represented by a diffusion process $x_t \in \mathbb{R}^n$ defined by stochastic differential equation (SDE)[47]. The forward diffusion process is characterized by the following equation:

$$dx = f(x,t)dt + g(t)dw \tag{1}$$

where $t \in [0,T]$ is a continuous index and $w \in \mathbb{R}^n$ is the standard Wiener process (a.k.a., Brownian motion). $f(x,t) \in \mathbb{R}^n$ is a vector-valued function called the drift coefficient, and $g(t) \in \mathbb{R}$ is a scalar function called the diffusion coefficient. Then, the corresponding backward diffusion process, or denoising process can also be defined by SDE[48]:

$$dx = [f(x,t) - g^2(t)\nabla_x logp_t(x)]dt + g(t)d\overline{w} \tag{2}$$

where $d\overline{w}$ is a standard Wiener process as continuous time t flows backwards from T to 0, and dt is an infinitesimal negative time step. $x_{t=0}$ here represents ground truth data, or the protein conformations, while $x_{t=T}$ is sampled from Gaussian distribution. Therefore, by solving the backward SDE process, we can sample diverse protein conformations that obey Boltzmann distribution from a Gaussian distribution. In equation 2, each term except the score of x, $\nabla_x logp_t(x)$, is solvable. Therefore, we only needed a score-matching network to fit $\nabla_x logp_t(x)$ at each time step to achieve the purpose of generating conformations.

Based on the above-mentioned standard formulation of SGM, we further required that the entire conformation generation process should be SE(3)-equivariant, i.e., diffusion process and network transformation were not sensitive to global rotation and translation of protein structures. SE(3)-equivariance can be described by the following equation:

$$\mathcal{F} \circ \rho(x) = \rho \circ \mathcal{F}(x) \tag{3}$$

where $\mathcal{F}$ denotes data transformations like network prediction and diffusion process, ρ represents global rotation and translation. Typically, a protein conformation x is characterized by Cartesian coordinates $c_i \in \mathbb{R}^3, 1 \leq i \leq N$, where N denotes number of atoms. However, data transformations on Cartesian coordinates is computationally intensive and does not easily satisfy SE(3)-equivariance. Therefore, we adopted backbone frame parametrization, representing the protein conformation x as $T_j \coloneqq [R_j, v_j], 1 \leq j \leq n$, where n denotes the number of residues. Each backbone frame includes a 3×3 rotation matrix $R_j \in SO(3)$ and a translation vector $v_j \in \mathbb{R}^3$. A frame T_j can represent Euclidean transformation for each atom in residue j from local coordinates c_{local} to global coordinates c_{global} as $c_{global} = T_i \circ c_{local} \coloneqq R_i\, c_{local} + v_i$. Following the approach described in FrameDiff[43], we independently handle the rotation matrix R_j on the $SO(3)$ manifold and the translation vector v_j in $\mathbb{R}^3$ during the diffusion process, formulated as follows:

$$dT_t = \left[0, -\frac{1}{2}\beta(t)Pv_t\right]dt + \left[\sqrt{\frac{d}{dt}\sigma^2(t)}dw^{(SO(3))}, \sqrt{\beta(t)}Pdw^{(\mathbb{R}^3)}\right] \quad (4)$$

Where $\beta(t)$ and $\sigma(t)$ control the scale of noise during the diffusion process, $w^{\mathcal{M}}$ denotes Brownian motion defined on manifold $\mathcal{M}$, and $P: \mathbb{R}^{3n} \rightarrow \mathbb{R}^{3n}$ is used for removing the center of mass. During the forward diffusion or noising process, the addition of noise on rotation matrices is determined by the noise kernel $p_{t|0}(R_t|R_0)$, which is obtained from an isotropic Gaussian distribution on the $SO(3)$ manifold. This distribution is formulated as:

$$\mathcal{IG}_{SO(3)}(R_t; R_0, \sigma^2) = f(\omega_{t|0}) \coloneqq \frac{1-\cos(\omega_{t|0})}{\pi}\sum_{l=0}^{\infty}(2l+1)e^{-l(l+1)\sigma^2\frac{\sin((l+0.5)\omega_{t|0})}{\sin(0.5\omega_{t|0})}} \quad (5)$$

Here, $\omega_{t|0} = Axis_angle(R_0^T R_t)$ is the axis-angle transformed representation of the composed rotation matrix $R_0^T R_t$. As for the translation vector, its noise addition process follows an Ornstein-Uhlenbeck process, also known as VP-SDE. The noise kernel for the translation vector is relatively straightforward, converging ultimately to $\mathcal{N}(0, I)$ as shown in the following equation.

$$p_{t|0}(v_t|v_0) = \mathcal{N}\left(v_t; v_0 e^{-\frac{1}{2}\int_0^t \beta(s)ds}, I - Ie^{\int_0^t \beta(s)ds}\right) \quad (6)$$

Network Design and Training

To achieve the goal of predicting conformational ensembles from sequence, we devised a sequence-conditioned score-matching model $s_\theta(x_t, x, seq)$ for denoising processes. As the diffusion process is designed to be strictly SE(3)-equivariant, network transformations should evidently preserve this property. Therefore, we adopted a variant of structure module from AlphaFold2 (Figure 1A) for updating backbone frames[22,49]. Here, we employed the Point Invariant Attention (IPA) mechanism to capture interactions and relationships between nearby residues, followed by a Transformer to learn global features and long-range interactions. This architecture has been demonstrated in previous research to promote training and generation of high-quality protein conformations. The aforementioned network design requires three inputs foreach layer: a one-dimensional vector representation s_l, pairwise feature representation z_l, and the set of rotation and translation updates T_l. We utilized ESM2-650M to extract protein sequence features, concatenated with residue position encoding and time encoding represented by trigonometric functions, to form the initial one-dimensional vector representation s_0. Pairwise feature representation z_0 is derived from s_0 based on relative positional encoding. After each layer of IPA-Transformer transformation, we updated the one-dimensional vector representation through a fully connected network and subsequently updated the pairwise features *via* cross product.

The objective of score-matching networks differs from conventional neural network training goals. It does not aim to fit protein conformations directly but rather the scores of data perturbed to a certain degree, i.e., fitting the distribution of perturbed data. To measure how well the predicted scores fit actual distribution, we computed the DSM loss as follows:

$$\mathcal{L}_{dsm} = \mathbb{E}_{t\in[0,T]}\left\{\lambda(t)\mathbb{E}_{T_0}\mathbb{E}_{T_t|T_0}\left[\left|\left|s_\theta(T_t, t) - \nabla_{T_t} log p_{t|0}(T_t|T_0)\right|\right|^2\right]\right\} \quad (7)$$

To ensure that the DSM loss at all time steps t results in a perfect fit score of 1, ensuring equal contribution of each time step to the loss function, we set the weights as

follows:

$$\lambda(t) = \frac{1}{\mathbb{E}\left[\left\|\nabla_{T_t} log p_{t|0}(T_t|T_0)\right\|\right]} \tag{8}$$

Additionally, to ensure the model learns detailed features of protein structures, apart from the DSM loss on rotation and translation matrices, we also incorporate mean square error (MSE) supervision for the positions of backbone atoms and differences on the distance matrix for samples with fewer forward diffusion steps ($t < \frac{T}{4}$). Therefore, the complete network training loss function can be represented as:

$$\mathcal{L} = \mathcal{L}_{dsm} + \omega_1 \mathcal{L}_{bb} + \omega_2 \mathcal{L}_{dist} \tag{9}$$

where $\omega_1 = \omega_2 = 0.25$ control the weight of conformation quality loss.

During the training process, with a maximum time step $T = 1.0$, we optimize the DenoisingIPA using an Adam optimizer with the learning rate of 10^{-4}[50]. For the translation vector part of the network training, we employ a linear noise strategy within the VP-SDE framework, while for the rotation matrix part, we use a logarithmic noise strategy within the VE-SDE framework, as shown in the following equation[47]:

$$\beta(t) = \beta_{min} + \frac{t}{T}(\beta_{max} - \beta_{min}),$$
$$\beta_{min} = 0.1, \beta_{max} = 20 \tag{10}$$

$$\sigma(t) = \log(te^{\sigma_{max}} + (T - t)e^{\sigma_{min}}),$$
$$\sigma_{min} = 0.1, \sigma_{max} = 1.5 \tag{11}$$

Implementation Details

The training of IDPFold consists of two stages: pre-training on experimental structures and fine-tuning on MD trajectories. In Table 2, we present the model hyperparameters utilized during the 2 training stages:

Table 2. Hyperparameters and training details of IDPFold

Hyperparameters	**Training Stage**	
	Training on Experimental Data	Training on MD Trajectories

Single Repr. Channel	256	
Pair Repr. Channel	128	
Hidden Channel	256	
IPA Layers	4	
Transformer Layers	2	
Transformer Heads	8	
Model Size	17.8 M parameters	
Learning Rate	10^{-4}	10^{-5}
Batch Size	8	32
Iterations	1.12 M	0.44 M
Time	~9 GPU days	~15 GPU days

Evaluation Metrics and Analysis Tools

We evaluated the quality of IDPFold-predicted IDP conformational ensembles from two perspectives: local features of the generated structures and global features of the conformational ensembles. For local features, biotite is used to compute inter-residue bond lengths and bond angles of all generated conformations, specifically, the $C-N$ bond lengths, $C_\alpha - C - N$ and $C - N - C_\alpha$ bond angles[51]. We also analyzed the distribution of backbone dihedral angles to assess the model's prediction accuracy regarding C_α chirality and secondary structure. Additionally, we used mdtraj to calculate scalar coupling between HN and H_α, employed SPARTA+ to compute C_α chemical shifts, and then performed ensemble averaging[52,53]. These analyses helped determine whether the local environment of the protein backbone aligns with experimental observations.

For global features, we calculated Rg of the generated conformations and compared the ensemble average with experiment results. Additionally, we derived C_α distance matrices to facilitate PCA dimensionality reduction. Through clustering the generated conformations and projecting them onto principal component space, we aim to discern the diversity of generated conformations and assess the model's learning efficacy in capturing the Boltzmann distribution information from MD trajectories in the training

dataset.

Due to the fact that most current work on generating protein conformational ensembles from sequences adopts a coarse-grained representation, where only C_α coordinates are generated, we first fixed the generated structures with pdbfixer and ran a 100-step energy minimization. During minimization we added restraints on C_α to make sure the optimized backbones did not differ from the original ones too much. This process cost about 6 seconds per conformation. Then we evaluated Validity and Fidelity and compared the performance of IDPFold with previous methods.

Validity assesses whether the generated conformations contain unreasonable C_α distances. In a protein structure, due to van der Waals interactions between atoms, C_α distances should not be too close. Additionally, because neighboring C_α in the protein backbone are connected by $C - N$ with specific bond lengths, C_α distances should not be excessively far apart. Therefore, following the approach of Str2Str, we defined a reasonable range for C_α distances as follows:

$$\delta_{vdw} < d_{C_\alpha^i, C_\alpha^{i+1}} < \delta_{bond} \tag{12}$$

where $\delta_{vdw} = 2 \times 1.7 - 0.4$ is defined as the sum of two C_α van der Waals radius minus an acceptable overlap distance of 0.4[54]. δ_{bond} is taken as the maximum C_α distance observed in MD trajectory of each test system. Since a reasonable range is defined, we could define validity as proportion of valid conformations. A higher validity indicates a lower probability of mis-estimated C_α distances, thus demonstrating better performance.

Fidelity refers to how well the model-generated ensembles match experimental observations. We selected Rg and C_α secondary chemical shifts as target physical quantities to measure the fidelity of models, calculating the errors between model-generated conformational ensembles and experimental observations using the following formula:

$$\varepsilon_{Rg} = \frac{\langle Rg \rangle^{pred} - \langle Rg \rangle^{exp}}{\langle Rg \rangle^{exp}} \tag{13}$$

$$MAE_{\delta C_\alpha} = \mathbb{E}(|\langle \delta C_\alpha \rangle^{pred} - \langle \delta C_\alpha \rangle^{exp}|) \tag{14}$$

where $\langle Rg \rangle$ denotes the ensemble average Rg, $\langle Rg \rangle^{pred}$ is the generated ensemble average Rg while $\langle Rg \rangle^{exp}$ is experimental observation. Similarly, $\langle \delta C_\alpha \rangle^{pred}$ and $\langle \delta C_\alpha \rangle^{exp}$ denotes ensemble average chemical shift and experimental observation respectively. Using the above definitions, we calculated ε_{Rg} and $MAE_{\delta C_\alpha}$, aiming for values close to zero for both. ε_{Rg} reflects the reasonableness of the model's generated conformations in terms of their compactness, while $MAE_{\delta C_\alpha}$ indicates how well the generated conformations conform to experimental observations on local secondary structures.

For evaluating the conformational ensembles generated by the models, all methods except AF-cluster generated 300 conformations for subsequent evaluation[55,56]. This number of conformations has been previously shown to adequately reflect the structural diversity of intrinsically disordered proteins. The number of conformations generated by AF-cluster is influenced by the number of clusters in the MSA, and we used its default MSA and clustering settings. All conformation generation was performed on NVIDIA A100 GPUs.

Acknowledgements

This work was supported by the Center for HPC at Shanghai Jiao Tong University, and the National Key Research and Development Program of China (2023YFF1205102 and 2020YFA0907700), the Fundamental Research Funds for the Central Universities (YG2023LC03), and the National Natural Science Foundation of China (21977068, 32171242 and T2222012).

Code Availability

The code of IDPFold is available at https://github.com/Junjie-Zhu/IDPFold.

References

1. Dunker, A. K. *et al.* The unfoldomics decade: an update on intrinsically disordered proteins. *BMC Genomics* **9**, S1 (2008).
2. Dunker, A. K., Brown, C. J., Lawson, J. D., Iakoucheva, L. M. & Obradović, Z. Intrinsic Disorder and Protein Function. *Biochemistry* **41**, 6573–6582 (2002).
3. Uversky, V. N., Oldfield, C. J. & Dunker, A. K. Intrinsically Disordered Proteins in Human Diseases: Introducing the D ^{2} Concept. *Annu. Rev. Biophys.* **37**, 215–246 (2008).
4. Iakoucheva, L. M., Brown, C. J., Lawson, J. D., Obradović, Z. & Dunker, A. K. Intrinsic Disorder in Cell-signaling and Cancer-associated Proteins. *J. Mol. Biol.* **323**, 573–584 (2002).
5. Zhao, R. *et al.* Analysis of p53-regulated gene expression patterns using oligonucleotide arrays. *Genes Dev.* **14**, 981–993 (2000).
6. Singleton, A. B. *et al.* α-Synuclein Locus Triplication Causes Parkinson's Disease. *Science* **302**, 841–841 (2003).
7. Casu, F., Duggan, B. M. & Hennig, M. The Arginine-Rich RNA-Binding Motif of HIV-1 Rev Is Intrinsically Disordered and Folds upon RRE Binding. *Biophys. J.* **105**, 1004–1017 (2013).
8. Fernández, A. & Scheraga, H. A. Insufficiently dehydrated hydrogen bonds as determinants of protein interactions. *Proc. Natl. Acad. Sci. U. S. A.* **100**, 113–118 (2003).
9. Chruszcz, M., Borek, D., Domagalski, M., Otwinowski, Z. & Minor, W. X-ray diffraction experiment – the last experiment in the structure elucidation process. *Adv. Protein Chem. Struct. Biol.* **77**, 23–40 (2009).
10. Dandey, V. P. *et al.* Time-resolved cryoEM using Spotiton. *Nat. Methods* **17**, 897–900 (2020).
11. McDermott, A. Structure and Dynamics of Membrane Proteins by Magic Angle Spinning Solid-State NMR. *Annu. Rev. Biophys.* **38**, 385–403 (2009).
12. Jensen, M. R., Zweckstetter, M., Huang, J. & Blackledge, M. Exploring Free-Energy Landscapes of Intrinsically Disordered Proteins at Atomic Resolution Using NMR Spectroscopy. *Chem. Rev.* **114**, 6632–6660 (2014).
13. Huang, J. *et al.* CHARMM36m: An Improved Force Field for Folded and Intrinsically Disordered Proteins. *Nat. Methods* **14**, 71–73 (2017).
14. Robustelli, P., Piana, S. & Shaw, D. E. Developing a molecular dynamics force field for both folded and disordered protein states. *Proc. Natl. Acad. Sci. U. S. A.* **115**, E4758–E4766 (2018).
15. Li, Z., Mu, J., Chen, J. & Chen, H.-F. Base-specific RNA force field improving the dynamics conformation of nucleotide. *Int. J. Biol. Macromol.* **222**, 680–690 (2022).
16. Zhang, Y., Liu, H., Yang, S., Luo, R. & Chen, H.-F. A Well-balanced Force Field ff03CMAP for Folded and Disordered Proteins. *J. Chem. Theory Comput.* **15**, 6769–6780 (2019).
17. Song, D. *et al.* ff14IDPs Force Field Improving the Conformation Sampling of Intrinsically Disordered Proteins. *Chem. Biol. Drug Des.* **89**, 5–15 (2017).
18. Song, D., Liu, H., Luo, R. & Chen, H.-F. Environment-Specific Force Field for Intrinsically Disordered and Ordered Proteins. *J. Chem. Inf. Model.* **60**, 2257–2267 (2020).
19. Zhong, B., Song, G. & Chen, H.-F. Balanced Force Field ff03CMAP Improving the Dynamics

Conformation Sampling of Phosphorylation Site. *Int. J. Mol. Sci.* **23**, 11285 (2022).
20. Mu, J., Pan, Z. & Chen, H.-F. Balanced Solvent Model for Intrinsically Disordered and Ordered Proteins. *J. Chem. Inf. Model.* **61**, 5141–5151 (2021).
21. Tesei, G. *et al.* Conformational ensembles of the human intrinsically disordered proteome. *Nature* **626**, 897–904 (2024).
22. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
23. Lin, Z. *et al.* Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **379**, 1123–1130 (2023).
24. Watson, J. L. *et al.* De novo design of protein structure and function with RFdiffusion. *Nature* **620**, 1089–1100 (2023).
25. Ingraham, J. B. *et al.* Illuminating protein space with a programmable generative model. *Nature* **623**, 1070–1078 (2023).
26. Berman, H. M. *et al.* The Protein Data Bank. *Nucleic Acids Res.* **28**, 235–242 (2000).
27. Lazar, T. *et al.* PED in 2021: a major update of the protein ensemble database for intrinsically disordered proteins. *Nucleic Acids Res.* **49**, D404–D411 (2020).
28. Riley, A. C., Ashlock, D. A. & Graether, S. P. The difficulty of aligning intrinsically disordered protein sequences as assessed by conservation and phylogeny. *PLOS ONE* **18**, e0288388 (2023).
29. Pan, Z., Mu, J. & Chen, H.-F. Balanced Three-Point Water Model OPC3-B for Intrinsically Disordered and Ordered Proteins. *J. Chem. Theory Comput.* **19**, 4837–4850 (2023).
30. Wayment-Steele, H. K. *et al.* Predicting multiple conformations via sequence clustering and AlphaFold2. *Nature* **625**, 832–839 (2024).
31. Jing, B., Berger, B. & Jaakkola, T. AlphaFold Meets Flow Matching for Generating Protein Ensembles. Preprint at https://doi.org/10.48550/arXiv.2402.04845 (2024).
32. Janson, G. & Feig, M. Transferable deep generative modeling of intrinsically disordered protein conformations. 2024.02.08.579522 Preprint at https://doi.org/10.1101/2024.02.08.579522 (2024).
33. Janson, G., Valdes-Garcia, G., Heo, L. & Feig, M. Direct generation of protein conformational ensembles via machine learning. *Nat. Commun.* **14**, 774 (2023).
34. Cao, F., Von Bülow, S., Tesei, G. & Lindorff-Larsen, K. A coarse-grained model for disordered and multi-domain proteins. Preprint at https://doi.org/10.1101/2024.02.03.578735 (2024).
35. Tesei, G. & Lindorff-Larsen, K. Improved predictions of phase behaviour of intrinsically disordered proteins by tuning the interaction range. *Open Res. Eur.* **2**, 94 (2023).
36. Uversky, V. N., Gillespie, J. R. & Fink, A. L. Why are “natively unfolded” proteins unstructured under physiologic conditions? *Proteins Struct. Funct. Bioinforma.* **41**, 415–427 (2000).
37. Firouzbakht, A. *et al.* HYPK: A marginally disordered protein sensitive to charge decoration. *Proc. Natl. Acad. Sci.* **121**, e2316408121 (2024).
38. Kobashigawa, Y. *et al.* Structural basis for the transforming activity of human cancer-related signaling adaptor protein CRK. *Nat. Struct. Mol. Biol.* **14**, 503–510 (2007).
39. Ota, T. *et al.* Complete sequencing and characterization of 21,243 full-length human cDNAs. *Nat. Genet.* **36**, 40–45 (2004).
40. Abramson, J. *et al.* Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature* 1–3 (2024) doi:10.1038/s41586-024-07487-w.
41. Raychaudhuri, S., Sinha, M., Mukhopadhyay, D. & Bhattacharyya, N. P. HYPK, a Huntingtin

interacting protein, reduces aggregates and apoptosis induced by N-terminal Huntingtin with 40 glutamines in Neuro2a cells and exhibits chaperone-like activity. *Hum. Mol. Genet.* **17**, 240–255 (2008).

42. Zheng, S. *et al.* Towards Predicting Equilibrium Distributions for Molecular Systems with Deep Learning. Preprint at https://doi.org/10.48550/arXiv.2306.05445 (2023).
43. Yim, J. *et al.* SE(3) diffusion model with application to protein backbone generation. Preprint at http://arxiv.org/abs/2302.02277 (2023).
44. Noé, F., Olsson, S., Köhler, J. & Wu, H. Boltzmann generators: Sampling equilibrium states of many-body systems with deep learning. *Science* **365**, eaaw1147 (2019).
45. Yang, J. *et al.* Improved protein structure prediction using predicted interresidue orientations. *Proc. Natl. Acad. Sci.* **117**, 1496–1503 (2020).
46. Maier, J. A. *et al.* ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. *J. Chem. Theory Comput.* **11**, 3696–3713 (2015).
47. Song, Y. *et al.* Score-Based Generative Modeling through Stochastic Differential Equations. Preprint at http://arxiv.org/abs/2011.13456 (2021).
48. Anderson, B. D. O. Reverse-time diffusion equation models. *Stoch. Process. Their Appl.* **12**, 313–326 (1982).
49. Lu, J., Zhong, B., Zhang, Z. & Tang, J. Str2Str: A Score-based Framework for Zero-shot Protein Conformation Sampling. in (2023).
50. Kingma, D. P. & Ba, J. Adam: A Method for Stochastic Optimization. Preprint at http://arxiv.org/abs/1412.6980 (2017).
51. Kunzmann, P. & Hamacher, K. Biotite: a unifying open source computational biology framework in Python. *BMC Bioinformatics* **19**, 346 (2018).
52. McGibbon, R. T. *et al.* MDTraj: A Modern Open Library for the Analysis of Molecular Dynamics Trajectories. *Biophys. J.* **109**, 1528–1532 (2015).
53. Shen, Y. & Bax, A. SPARTA+: a modest improvement in empirical NMR chemical shift prediction by means of an artificial neural network. *J. Biomol. NMR* **48**, 13–22 (2010).
54. Ramachandran, S., Kota, P., Ding, F. & Dokholyan, N. V. Automated Minimization of Steric Clashes in Protein Structures. *Proteins* **79**, 261–270 (2011).
55. Daughdrill, G. W. *et al.* Understanding the Structural Ensembles of a Highly Extended Disordered Protein. *Mol. Biosyst.* **8**, 308–319 (2012).
56. Schwalbe, M. *et al.* Predictive Atomic Resolution Descriptions of Intrinsically Disordered hTau40 and α-Synuclein in Solution from NMR and Small Angle Scattering. *Structure* **22**, 238–249 (2014).